Received: 9 April 2019 | Revised: 14 November 2019 | Accepted: 15 November 2019

DOI: 10.1002/kin.21335

WILEY

ARTICLE

Automatic kinetic model generation and selection based on concentration versus time curves

Tibor Nagy[1,2] | **János Tóth**[2,3] | **Tamás Ladics**[4]

[1]Institute of Materials and Environmental Chemistry, Research Centre for Natural Sciences, Hungarian Academy of Sciences, Budapest, Hungary

[2]Laboratory for Chemical Kinetics, Eötvös Loránd University, Budapest, Hungary

[3]Department of Mathematical Analysis, Budapest University of Technology and Economics, Budapest, Hungary

[4]Department of Science and Engineering, John von Neumann University, Kecskemét, Hungary

Correspondence
Tibor Nagy, Institute of Materials and Environmental Chemistry, Research Centre for Natural Sciences, Hungarian Academy of Sciences, 1117 Budapest, Hungary.
Email: nagy.tibor@ttk.mta.hu

Dedicated to Professors László Lovász and József Pálinkás.

Funding information
Magyar Tudományos Akadémia, Grant/Award Number: BO/00279/16/7; National Research, Development and Innovation Office, Grant/Award Numbers: PD 120776, SNN 125739

Abstract

The goal of the paper is to automatize the construction and parameterization of kinetic reaction mechanisms that can describe a set of experimentally measured concentration versus time curves. Using the framework and theorems of formal reaction kinetics, first, we build a set of possible mechanisms with a given number of measured and unmeasured (real or fictitious) species and reaction steps that fulfill some chemically reasonable requirements. Then we fit all the corresponding mass-action kinetic models and offer the best one to the chemist to help explain the underlying chemical phenomenon or to use it for predictions. We demonstrate the use of the method via two simple examples: on an artificial, simulated set of data and on a small real-life data set. The method can also be used to do a kind of lumping to generate a model that can reproduce the simulation results of a detailed mechanism with less species and thereby can largely accelerate spatially inhomogeneous simulations.

KEYWORDS

automatically generated models, fitting rate coefficients, inverse problem of reaction kinetics, model parameterization, reaction networks, reduced models

1 | INTRODUCTION

Our goal is to help solve an inverse problem of reaction kinetics: We try to build "reaction mechanism templates," candidate sets of reactions which are able to serve as models for given (measured or simulated with a detailed model) concentration versus time curves. This we do by constructing a large set of reaction mechanisms then by discarding those which do not fulfill some chemically relevant restrictions. To put it in another way, we provide a set of possible models automatically for the chemist, and say, which of them fits best to the experimental data or the simulation results of a detailed model.

Automatic mechanism and kinetic model generation protocols and codes have been developed and widely used in the fields of biochemistry,[1] organic chemistry,[2] atmospheric chemistry,[3,4] and combustion chemistry.[5,6] Recently, a new field of application is emerging in reactive molecular

dynamics,[7] where noisy concentration profiles are generated as a result of atomistic simulations, and one needs to deduce the relevant microkinetic mechanism. These approaches are all based on detailed chemical knowledge upon the possible type of species and reactions and estimate rate coefficients by applying rate theories or analogues. In contrast, we propose a procedure for model development that does not use any specific microscopic chemical information on the system, and the rate parameterization is also done fully empirically, by fitting the model results to previously determined concentration data.

Fitting kinetic models requires nonlinear parameter estimation, which has a rich literature. Reference 8 is a classic book, and Refs. 9 and 10 are recent monographs. The first pioneering works aimed at the special problem of estimating reaction rate coefficients in detailed reaction mechanisms were published in combustion chemistry.[11,12] Recent works on kinetic parameter estimations are Refs. 13 and 14 and Refs. 15,16, and 17. The latter three papers describe an efficient way for determining Arrhenius parameters, which define the temperature dependence of the rate coefficient.

Thus, given are concentrations as functions of time, and our task is to find the best-fitting *model* (or models), a *kinetic reaction mechanism*, that is a set of reaction steps (also called complex chemical reaction, reaction network, or simply reaction) endowed with mass action type kinetics and with appropriate reaction rate coefficients from the set of *all* possible models. We show how the method works by testing on two examples: on noisy *simulated* data and also a set of real-life data: measurements on the salicylic acid transport, which can be formally treated as a reaction kinetic problem. In the former example, we also demonstrate a type of indistinguishability of kinetic models[18–21] by showing that multiple mechanisms can perform equally well if the noise on the simulated data increased sufficiently. In the second one, not all the concentrations are measured and it serves as an example for the introduction of fictitious species.

The structure of our paper is as follows: Section 2 offers a possible scenario to construct candidate reaction mechanisms, which are to describe the concentration versus time curves. Section 3 shows the applicability of the method on two examples. The last section is about possible extensions and formulates some open problems related to the applicability of the method.

2 | A SET OF CHEMICALLY REASONABLE RESTRICTIONS

Basic concepts of formal reaction kinetics are presented here briefly; for the formal and detailed expansion, we propose the use of, for example, Ref. 22 or 23. Notations for the number of species, complexes, steps, etc. are summarized in Table 1 to aid the reader.

TABLE 1 Notations

Notation in formal kinetics	Meaning
$X, Y, Z, \ldots, X_m$	Species
L	Number of linkage classes
M	Number of species
N	Number of complexes
P	Number of reaction steps = 2 rev. pairs + irreversible steps
R	Number of reversible pairs of reactions
S	Number of independent reactions = rank $\boldsymbol{\gamma}$
$M_{\mathrm{rel}}(X_m)$	Relative molecular mass of species X_m
α_{mr}	Stoichiometric coefficient of species X_m on the left side of reaction step r
β_{mr}	Stoichiometric coefficient of species X_m on the right side of reaction step r
γ_{mr}	$\beta_{mr} - \alpha_{mr}$
k_r, k_r^+	Rate coefficient of the forward direction of reaction step r
k_{-r}, k_r^-	Rate coefficient of the backward direction of reaction step r
K_r	Equilibrium constant of reversible reaction step r
c_m	Concentration of species X_m

2.1 | Species

First, we have to fix M, the number of species. In our illustrating examples, it will usually be two or three; in applications, this may be equal to the number of measured concentrations or can be more by introducing further unmeasured (real or fictitious) species.

2.2 | Complexes

In formal reaction kinetics, the linear combinations of species with stoichiometric coefficients (or stoichiometric numbers) on the sides of the reactions are called *complexes*, a slightly unfortunate name, because this word is used in chemistry with a completely different meaning. Their number is usually denoted by N. The stoichiometric coefficient of the mth species (X_m) on the left side of the rth reaction is denoted by α_{mr}, that on the right side by β_{mr}, thus the general form of the rth reversible reaction step in our mechanism is

$$\sum_{m=1}^{M} \alpha_{mr} X_m \rightleftharpoons \sum_{m=1}^{M} \beta_{mr} X_m. \tag{1}$$

2.2.1 | Mass conservation

As a consequence of mass conservation $M = 1$ is immediately excluded, because, for example, the reaction step $X \rightleftharpoons 2X$ is not allowed. Also excluded is the *empty complex* 0, as it immediately leads to mass destruction or mass

creation. Still, in other contexts, the empty complex may be useful either to describe the sticking of a species to the wall or its departure in any other way, or to express an inflow or outflow in formal reactions, for example, $0 \longrightarrow X$ or $X + Y \longrightarrow 0$.

2.2.2 | Short complexes

In most cases reaction steps containing complexes longer than two, are not allowed; here we follow this practice. In other words, we only allow *short complexes*, complexes of the form in which no more than two species take part. In the case when $M = 2$, we have thus X,Y,2X,2Y,X + Y (zero complex excluded from the beginning). Their number in the general case, as it can be immediately seen, is

$$\overline{N}(M) = M + M + \binom{M}{2} = \frac{M(M+3)}{2}. \tag{2}$$

Let us mention in passing that the concept of short complexes proved to be really useful when the dynamic behavior of small reaction mechanisms are investigated.[24]

If a formal reaction represents an elementary chemical reaction then the length of the left-side complex (ie, $\Sigma_{m=1}^{M}\alpha_{mr}$) is equal to the molecularity of the reaction. Consequently, a reaction mechanism built up from reversible steps of short complexes (ie, $1 \le \Sigma_{m=1}^{M}\alpha_{mr} \le 2$ and $1 \le \Sigma_{m=1}^{M}\beta_{mr} \le 2$) can describe reacting systems involving only unimolecular and bimolecular elementary reactions. However, our present restriction does not mean that termolecular reactions are unimportant. In the gas phase, reactions like

$$2NO + O_2 \longrightarrow 2NO_2$$

are quite common (see, eg, Ref. 25). To take these into consideration one should also include 3X, 2X + Y,…, X + Y + Z, the total number of which is

$$M + M(M-1) + \binom{M}{3} = \frac{M(M+1)(M+2)}{6} = \binom{M+2}{3}. \tag{3}$$

Note that the result is actually the formula for the number of three-combinations of M species when repetition is allowed.

2.3 | Reaction steps

Considering the transformation of two complexes into each other gives a reaction step. In formal reaction kinetics, no duplicate of a reaction step is allowed as repeated steps can be combined into a single step whose rate coefficient is the sum of individual rate coefficients. Furthermore, if both directions of a reaction are present in the mechanism then they are always written as a single reversible step, and thus irreversible steps cannot have a reverse pair in the mechanism.

2.3.1 | Reversibility

In chemistry, all elementary reactions are strictly reversible due to microscopic reversibility. However, in practical applications at certain conditions (eg, concentrations, temperature, and pressure), the rate of the forward or the backward reaction can become negligible, thus this direction can be omitted from the model without inducing significant error and the retained direction can be considered as an irreversible step.

Accordingly, in widely accepted detailed combustion models some irreversible steps are also used, for example,

- in hydrogen combustion[26]

$$H_2O_2 + O \longrightarrow HO_2 + OH,$$

- in carbon monoxide combustion[26]

$$HCO + HO_2 \longrightarrow CO_2 + H + OH.$$

Nevertheless, during model construction we assume that the reaction steps are reversible. The advantage of this approach during model building is that the number of possible reaction steps (and that of models) is less than when both reversible and irreversible reactions are considered; furthermore, the omission of one of the directions can be investigated a posteriori using mechanism reduction techniques (see, eg, Refs. 27,28, and 16). Leaving out either the forward direction or the backward direction will increase the number of possible steps by a factor of three (ie, leave out forward or backward step or none) per each reversible step.

2.3.2 | Macroscopic chemical change

Certainly there is no reason to include reaction steps like $X \rightleftharpoons X$, $2X \rightleftharpoons 2X$, $X + Y \rightleftharpoons X + Y$, which do not affect concentrations (ie, make no macroscopic change) despite that they may take place microscopically.

2.3.3 | Mass conservation

From the chemical point of view, it is quite reasonable to require that the total mass in both sides of reaction (1) be the same. This can be formalized in such a way that using the relative molecular masses $M_{\mathrm{rel}}(X_m)$ of species X_m the following equality should hold for the rth reaction:

$$\sum_{m=1}^{M} \alpha_{mr} M_{\mathrm{rel}}(X_m) = \sum_{m=1}^{M} \beta_{mr} M_{\mathrm{rel}}(X_m). \tag{4}$$

Mass conservation is a consequence of a fundamental property of chemical reactions, the law of atomic balance, which requires the conservation of the number of various atomic nuclei and charge (ie, the number of electrons, thus their masses, too) in a chemical reaction.

In systems where either the molar number of species matters or mass transport is modeled, and chemical reactions significantly modify the mole fraction of species (eg, changes

are larger than 0.01); mass conservation in each reaction steps has to be strictly fulfilled for accurate simulation results. In most applications, one can neglect this restriction; however, for example, in combustion modeling it is usually needed. We shall also consider mass conservation here to reduce the number of possible reactions and mechanisms (see later).

Why should one consider reactions that are not mass conserving? In heterogeneous systems in the presence of a wall, an adsorbing material or a heterogeneous catalyst, adsorption, desorption of a bulk species, or its reaction with an adsorbed species can take place, which all lead to mass-violating bulk-phase reactions like $X \longrightarrow 0, 0 \longrightarrow X$ or $X \longrightarrow Y$, where $M_{rel}(X) \neq M_{rel}(Y)$, respectively. It is quite a common modeling tool to use steps like $0 \longrightarrow X$ to describe inflow, steps like $X \longrightarrow 0$ to represent outflow, or abbreviate a step like $A + X \longrightarrow 2X$ as $X \longrightarrow 2X$ if the concentration of the species A is so large that it practically does not change during the time we are interested in. For example, this is the case in atmospheric chemistry, where models often contain reaction steps of trace gases that produce major components of air (eg, N_2, O_2, CO_2 in tropospheric chemistry models) and the latter are simply omitted from the reaction as only negligible change in their concentration is induced by the step. Steps like $X \rightleftharpoons 2X$, $X \rightleftharpoons X + Y$, which obviously violate the law of mass conservation, may be quite useful in model construction in chemical kinetics. For example, the Lotka-Volterra reaction, which can be used for *approximately* describing oscillations in cold flames,[29] contains steps like $X \longrightarrow 2X$ and $Y \longrightarrow 0$.

Still, in the present example we shall take a more standard (conservative, if you wish) point of view, and we discard steps like $0 \rightleftharpoons X$, $0 \rightleftharpoons 2X$, $0 \rightleftharpoons X + Y$, $X \rightleftharpoons 2X$, $X \rightleftharpoons X + Y$, as they obviously violate the law of mass conservation. The number of possible reaction steps is less in this case, and one still has the a posteriori opportunity to remove bulk species with constant concentration (ie, incorporate their concentration into the rate coefficient) or to remove those product species from a step whose concentration is affected negligibly by the step using mechanism reduction methods.[16,27,28]

Let us mention in passing that chemical reactions are not perfectly mass-conserving. and there is a practically undetectable mass change arising in them due to the accompanied energy change as known from relativity theory (ie, mass-energy equivalence formula of Einstein[30]). On the other hand, while the kinetics of nuclear reactions (eg, radioactive decay, chain reactions) can also be formally described as that of chemical reactions, they do not fulfilled atomic balance and even mass conservation can break down badly in them due to the large energy changes.

2.3.4 | Symmetric self-reactions

In models, it is rare to see steps like $2X \rightleftharpoons 2Y$. However, studying the literature of atmospheric chemistry one can find similar steps: the self-reaction of peroxy radicals, like

$$2\,CH_3O_2 \rightleftharpoons 2\,CH_3O\,(+O_2),$$

or

$$2\,CH_3CH_2O_2 \rightleftharpoons 2\,CH_3CH_2O\,(+O_2),$$

(see Refs. 31 and 32). Both steps are of the form $2X \rightleftharpoons 2Y\,(+Z)$ and are mass conserving—only if Z (ie, O_2) is taken into consideration. As discussed before, one can omit the forming O_2 in atmospheric (eg, tropospheric) chemistry as it is a major constituent of air and its concentration will be negligibly affected by this and similar transformations.

Both steps are of the form $2X \rightleftharpoons 2Y\,(+Z)$ (and are mass conserving—if O_2 is taken into consideration—as they fulfilled the law of atomic balance).

TABLE 2 Species, complexes, and reaction steps in the case $M = 2$

Species	X, Y	$M = 2$
Complexes	X, Y, 2X, 2Y, X + Y	$\overline{N}(2) = 5$
Reaction steps	$X \rightleftharpoons Y$, $X \rightleftharpoons 2Y$, $Y \rightleftharpoons 2X$, $2X \rightleftharpoons X + Y$, $2Y \rightleftharpoons X + Y$	$\overline{R}(2) = 5$

TABLE 3 Species, complexes, and reaction steps in the case $M = 3$

Species	X, Y, Z	$M = 3$
Complexes	X, Y, Z, 2X, 2Y, 2Z, X + Y, Y + Z, Z + X	$\overline{N}(3) = 9$
Reaction steps	$X \rightleftharpoons Y$, $X \rightleftharpoons Z$, $Y \rightleftharpoons Z$, $X \rightleftharpoons 2Y$, $X \rightleftharpoons 2Z$, $Y \rightleftharpoons 2X$, $Y \rightleftharpoons 2Z$, $Z \rightleftharpoons 2X$, $Z \rightleftharpoons 2Y$, $X \rightleftharpoons Y + Z$, $Y \rightleftharpoons X + Z$, $Z \rightleftharpoons X + Y$, $2X \rightleftharpoons X + Y$, $2X \rightleftharpoons X + Z$, $2X \rightleftharpoons Y + Z$, $2Y \rightleftharpoons X + Y$, $2Y \rightleftharpoons X + Z$, $2Y \rightleftharpoons Y + Z$, $2Z \rightleftharpoons X + Y$, $2Z \rightleftharpoons X + Z$, $2Z \rightleftharpoons Y + Z$, $X + Y \rightleftharpoons X + Z$, $X + Y \rightleftharpoons Y + Z$, $X + Z \rightleftharpoons Y + Z$	$\overline{R}(3) = 24$

2.3.5 | Reactions passing all the criteria

The species, complexes, and reaction steps for $M = 2$ and $M = 3$ are shown in Tables 2 and 3, respectively.

Now the interesting (from the combinatorial point of view) question arises: What is the number of reactions fulfilling the requirements formulated above? Beyond combinatorics, the formula is of practical relevance too: it gives us a hint if it is possible to deal with all the systems constructed in this way within a tolerable time. We enumerated the steps from $M = 1$ up to $M = 20$ and have found the following cardinal-

TABLE 4 Number of reaction steps of different type

No.	Type of step	$M =$	2	3	4
1	$X \rightleftharpoons Y$	$\frac{M(M-1)}{2}$	1	3	6
2	$X \rightleftharpoons 2Y$	$M(M-1)$	2	6	12
3	$2X \rightleftharpoons X+Y$	$M(M-1)$	2	6	12
4	$X+Y \rightleftharpoons Z$	$\frac{M(M-1)}{2}(M-2)$	0	3	12
5	$X+Y \rightleftharpoons 2Z$	$\frac{M(M-1)}{2}(M-2)$	0	3	12
6	$X+Y \rightleftharpoons X+Z$	$\frac{M(M-1)}{2}(M-2)$	0	3	12
7	$X+Y \rightleftharpoons Z+A$	$\frac{1}{2}\frac{M(M-1)}{2}\frac{(M-2)(M-3)}{2}$	0	0	3
1-7	All types	$\frac{(M-1)M(M^2+7M+2)}{8}$	5	24	69

ities: 0, 5, 24, 69, 155, 300, 525, 854, … . How to learn if there is a certain regularity in the sequence? The best way is to go to The *On-Line Encyclopedia of Integer Sequences* initiated by Neil James Alexander Sloane[33] and ask if it contains our sequence. In this case, the answer was yes, and the formula $(M-1)M(M^2+7M+2)/8$ is provided to give the number of reaction steps of the given type. From the strict mathematical point of view, this statement is only a conjecture, but it can be rigorously proved, as well.

Statement 1. Suppose the number of species, M, is larger than one. Then, the number of reversible, mass conserving reaction steps excluding steps of the form

$$X \rightleftharpoons X, 2X \rightleftharpoons 2X \text{ and } 2X \rightleftharpoons 2Y$$

is

$$\overline{R}(M) := \frac{(M-1)M(M^2+7M+2)}{8}. \tag{5}$$

Proof. The formula is proved by determining the number of different types of steps that can be constructed using combinatorics and summing them up. Table 4 shows the seven types of reactions that can be stated with short complexes, and it also contains their number in the general case and for $M = 2, 3, 4$. The total number of possibilities equals the empirically determined formula, which proves the statement. □

2.4 | Reaction mechanisms

A set of reaction steps is usually called a (reaction) *mechanism*. If rate coefficients of the steps are also provided, it is called a kinetic reaction mechanism. In formal reaction kinetics, alternative names (complex chemical) *reaction*, or *reaction network* are also used.

2.4.1 | Mechanisms containing exactly M species

We may start from three species (eg, X, Y, Z) and generate all possible models by taking all combination of the possible reactions, then we may arrive at a mechanism with only two species, X and Y, as, for example,

$$X \rightleftharpoons Y, X \rightleftharpoons 2Y, Y \rightleftharpoons 2X.$$

It would be desirable to exclude such cases.

TABLE 5 Number of mechanisms containing different number of reaction steps

Number of species	Number of steps (R)						
	1	2	3	4	5	6	≥ 6
$M \le 3$	24	276	2024	10 626	42 504	134 596	$\binom{\overline{R}(3)}{R}$
$M = 3$	9	246	1994	10 611	42 501	134 596	$\tilde{R}(3)$

Let $\tilde{R}(M, R)$ denote the number of mechanisms having exactly R reactions and exactly M species. Suppose we have $M = 3$ species, then the number of mechanisms with at most three species and with exactly three species, and having various number (up to six) of reactions are shown in Table 5.

In the case when we have at most three species, it may happen that we only have two, thus the number of mechanisms with exactly three species is $\tilde{R}(3, R) = \binom{\overline{R}(3)}{R} - \binom{\overline{R}(2)}{R}\binom{3}{2}$, because one can select any two of the three species in $\binom{3}{2}$ different ways. The last and following elements (ie, $R = 6$ and $R > 6$) in the two rows of the table are equal, as one can only have five reaction steps with two species ($\overline{R}(2) = 5$; see Table 2). Let us formulate the corresponding obvious and general statement.

Statement 2. Suppose the number of species, M, is larger than one. Then, the number of mechanisms that consist of R reversible, mass conserving reaction steps, excluding steps of the form $X \rightleftharpoons X$, $2X \rightleftharpoons 2X$, and $2X \rightleftharpoons 2Y$, and contain exactly M species is

$$\begin{aligned}\tilde{R}(M, R) &:= \binom{\overline{R}(M)}{R} - \sum_{i=2}^{M-1} \tilde{R}(i, R)\binom{M}{i} \\ &= \binom{\overline{R}(M)}{R} - \binom{\overline{R}(M-1)}{R}\binom{M}{M-1} \\ &\quad + \binom{\overline{R}(M-2)}{R}\binom{M}{M-2} - \cdots \\ &= \sum_{i=0}^{M-2} (-1)^i \binom{\overline{R}(M-i)}{R}\binom{M}{M-i}. \end{aligned} \tag{6}$$

Consequently, in case $R > \overline{R}(M-i)$ one only has the preceding terms.

Proof. The first expression counts the mechanisms by excluding cases with exactly $M-1, M-2, \ldots$ species from all possible models with at most M species. The second expression gives the number of such models with a series of exclusions and inclusions based on the sieve formula (see, eg,

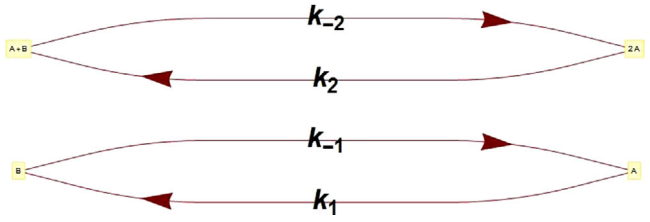

FIGURE 1 The Wegscheider mechanism [Color figure can be viewed at wileyonlinelibrary.com]

Ref. 34, Section 2). To understand the need for terms beyond the second, take the case of $M = 4$: the number of models with at most four species, $\{X, Y, Z, V\}$, has to be reduced by the number of models with at most three species: $\{X, Y, Z\}$, $\{X, Y, V\}$, $\{X, Z, V\}$, and $\{Y, Z, V\}$. However, when we do this, each two-species subset of them is actually excluded twice (eg, both $\{X, Y, Z\}$ and $\{X, Y, V\}$ have subset $\{X, Y\}$) thus they have to be included once again. For larger M (> 4), multiple inclusions can occur, which has to be corrected with exclusions and so on, leading to a series of terms with alternating signs, which is finally given in a compact sum form. □

2.4.2 | Mass conservation

Even if the steps are mass conserving, the mechanism may not be as example $\{X \rightleftharpoons Y, X \rightleftharpoons 2Y\}$ shows.

Definition 1. A reaction mechanism consisting of R reaction steps (see Equation 1; $r = 1, \ldots, R$) is said to be *mass conserving* if there exists a positive relative mass vector $\mathbf{M}_{\text{rel}} = (M_{\text{rel}}(X_1), \ldots, M_{\text{rel}}(X_M),)$ so that Equation 4 holds for all steps, which can be written shortly as a system of homogeneous linear equations for $\mathbf{M}_{\text{rel}}$:

$$\mathbf{0} = \mathbf{M}_{\text{rel}}(\boldsymbol{\beta} - \boldsymbol{\alpha}) = \mathbf{M}_{\text{rel}}\boldsymbol{\gamma} \tag{7}$$

with stoichiometric coefficient matrices $\boldsymbol{\alpha} := (\alpha_{mr})$, $\boldsymbol{\beta} := (\beta_{mr})$, and $\boldsymbol{\gamma} := \boldsymbol{\beta} - \boldsymbol{\alpha}$.

To check this property is not a trivial problem, we do this using our program `ReactionKinetics` described in Chap. 4 of Ref. 22, where the reader can also find relevant references as well. It is important to point out that this is a formal requirement that can even be stated for reactions involving fictitious species.

2.4.3 | Detailed balancing

Following Section 7.8 of Ref. 22, we review the history of detailed balancing shortly.

After such men as Maxwell[35] and Boltzmann,[36] and before Einstein,[37] at the beginning of the twentieth century, it was Wegscheider[38] who constructed the reaction mechanism in Figure 1 to show that in a closed system in some cases the existence of a positive stationary state (ie, when all steady-state concentrations are positive) alone does not imply the equality of all the individual forward and backward reaction rates: A relation (in this case $k_{-1}k_2 = k_{-2}k_1$) should hold between the reaction rate coefficients to ensure this. Equalities of this kind will be called (and later exactly defined as) *spanning forest conditions* below. Let us emphasize that violation of this equality between the reaction rate coefficients does not exclude the existence of a positive stationary state; it can be shown to exist and be unique for all values of the reaction rate coefficients. (Problem 7.12 of Ref. 22 proves both statements.)

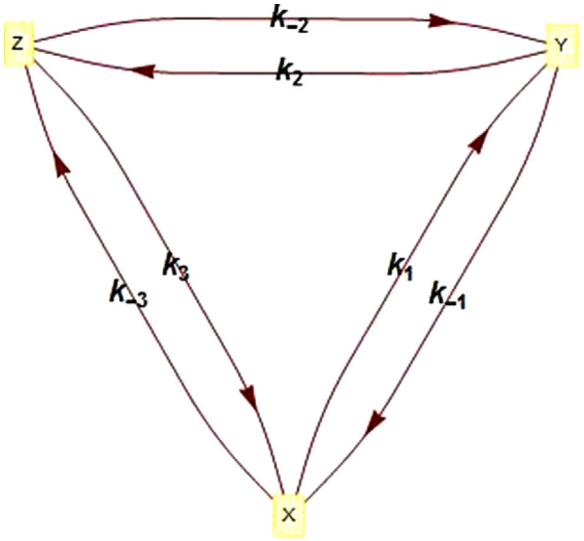

FIGURE 2 The reversible triangle reaction [Color figure can be viewed at wileyonlinelibrary.com]

Here we mention that Figures 1–4 show the Feinberg-Horn-Jackson graph (see, eg, Ref. 39 or Chap. 3 in Ref. 22) of three simple mechanisms, which is a directed graph with the complexes as vertices and with the reaction step arrows as directed edges. In this graph, each different complex of all the constituting reactions of the mechanism appears exactly once. The number of complexes is denoted by N, the number of connected components of the Feinberg-Horn-Jackson graph is L, whereas the number of independent reaction steps (the rank of $\boldsymbol{\gamma}$) is S.

A similar statement holds for the reversible triangle reaction in Figure 2. The necessary and sufficient condition for the existence of such a positive stationary state for which all the reaction steps have the same rate in the forward and backward direction (a detailed balanced stationary state) is now that the product of the reaction rate coefficients is the same if taken in either direction: $k_1k_2k_3 = k_{-1}k_{-2}k_{-3}$. Equalities of this kind will be called as *circuit conditions* below. Again, violation of this equality does not exclude the existence of a positive stationary state; it can again be shown to exist and be unique for all values of the reaction rate coefficients. (Problem 7.11 of the reference mentioned shows both statements.)

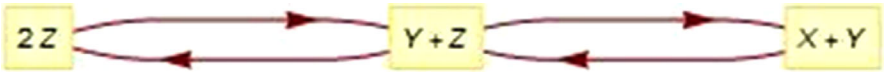

FIGURE 3 Unconditionally detailed balanced reaction mechanism: It is detailed balanced no matter what the values of the reaction rate coefficients are [Color figure can be viewed at wileyonlinelibrary.com]

These examples are qualitatively different from, for example, the simple one-step mechanism $X + Y \underset{k_{-1}}{\overset{k_1}{\rightleftharpoons}} Z$, which has the same stationary reaction rate in both directions no matter what the values of the reaction rate coefficients are (see Problem 7.10 in Ref. 22). To put it another way, this mechanism is *unconditionally detailed balanced* whereas the previous examples were *conditionally detailed balanced* (ie, being detailed balanced only if some equalities are fulfilled). A less trivial example is shown in Figure 3.

A quarter of a century after Wegscheider, Fowler and Milne[40] formulated a general principle in a very vague form called the *principle of detailed balance* stating that in real thermodynamic equilibrium, all the subprocesses should be in dynamic equilibrium separately in such a way that they do not stop but proceed with the same velocity in both directions. Obviously, this also means that time is reversible at equilibrium; that is why this property may also be called *microscopic reversibility*, although it may be appropriate to reserve this expression for a similar property of the stochastic model (see Chap. 10 of Ref. 22). A relatively complete summary of the early developments was given by Tolman.[41] The modern formulation of the principle accepted by IUPAC[42] essentially means the same (given that the principle of charity is applied when reading): "The principle of microscopic reversibility at equilibrium states that, in a system at equilibrium, any molecular process and the reverse of that process occur, on the average, at the same rate."

Now we give a precise formulation of the concept in such a way that at a detailed balanced stationary point (which can only exist in a reversible reaction), the forward and reverse reactions of each reversible pair proceed with the same rate.

The reaction mechanism we study here consists of R reversible pairs of reaction steps like Equation 1 and their usual induced kinetic differential equations assuming mass action type kinetics (and disregarding the change of temperature, pressure, and reaction volume) is

$$\dot{c}_m = \sum_{r=1}^{R} (\beta_{mr} - \alpha_{mr}) \left(k_r^+ \prod_{p=1}^{M} c_p^{\alpha_{pr}} - k_r^- \prod_{p=1}^{M} c_p^{\beta_{pr}} \right), \quad (8)$$

where $c_m(t) := [X_m](t)$ is the concentration of species X_m. (Note that P denotes the total number of reaction steps and we reserve notation R for the half of the number of reaction steps, or the number of reversible pairs. Thus, in the case of reversible steps $P = 2R$.) Shortly,

$$\dot{\mathbf{c}} = \boldsymbol{\gamma}(\mathbf{k}^+ \odot \mathbf{c}^{\boldsymbol{\alpha}} - \mathbf{k}^- \odot \mathbf{c}^{\boldsymbol{\beta}}) \quad (9)$$

with notation $(\mathbf{c}^{\boldsymbol{\alpha}})_r := \prod_{p=1}^{M} c_p^{\alpha_{pr}}$, and with the componentwise (or Schur) product $\odot$ of vectors: $(\mathbf{k}^+ \odot \mathbf{c}^{\boldsymbol{\alpha}})_r = k_r^+ (\mathbf{c}^{\boldsymbol{\alpha}})_r$. Here the positive numbers $k_r^{\pm}$ are the *reaction rate coefficients*; the vectors formed from them are $\mathbf{k}^{\pm}$. The reaction is detailed balanced at the positive stationary concentration $\mathbf{c}_*$ if all the steps proceed with the same rate in both directions, or, to put it another way

$$\boldsymbol{\gamma}(\mathbf{k}^+ \odot \mathbf{c}_*^{\boldsymbol{\alpha}} - \mathbf{k}^- \odot \mathbf{c}_*^{\boldsymbol{\beta}}) = \mathbf{0} \text{ implies} \quad (10)$$

$$\mathbf{k}^+ \odot \mathbf{c}_*^{\boldsymbol{\alpha}} = \mathbf{k}^- \odot \mathbf{c}_*^{\boldsymbol{\beta}} \quad \text{or} \quad \boldsymbol{\gamma}^\top \log(\mathbf{c}_*) = \log(\mathbf{K}), \quad (11)$$

where $\mathbf{K} := \frac{\mathbf{k}^+}{\mathbf{k}^-}$, which is also evaluated componentwise.

Detailed balance may hold

- at any (positive) values of the reaction rate coefficients (unconditionally detailed balanced) or
- only if the values of the rate coefficients fulfilled certain conditions—for example, circuit or spanning tree conditions—(conditionally detailed balanced).

What are the necessary and sufficient conditions of this property? First we give an algebraic characterization that can be proved using Fredholm's alternative theorem.

Theorem 1 (see Refs. 22 and 43). *The reaction mechanism is detailed balanced, if and only if for all nonzero* vector *solutions to the system of linear equations* $\boldsymbol{\gamma}\mathbf{a} = \mathbf{0}$ *one has*

$$\mathbf{K}^{\mathbf{a}} = \mathbf{1}. \quad (12)$$

As the elements of $\boldsymbol{\gamma}$ are integers and the vectors $\mathbf{a}$ are solutions of a system of homogeneous linear equations, their coordinates can supposed to be integers. The theorem can be recomposed for the chemist in a rather intuitive form: A reaction mechanism is detailed balanced, if and only if for all independent linear combinations of reactions steps (ie, "$\sum_{r=1}^{R} a_r \cdot \{\sum_{m=1}^{M} \alpha_{mr} X_m \rightleftharpoons \sum_{m=1}^{M} \beta_{mr} X_m .\}$") that express no chemical change (ie, $\sum_{r=1}^{R} a_r \cdot (\beta_{mr} - \alpha_{mr}) = 0$ for $m = 1, \ldots, M$), the corresponding equilibrium constants, which are the products of equilibrium constants raised to the powers of the respective linear weights, are one (ie, $\prod_{r=1}^{R} K_r^{a_r} = 1$).

Next we cite a pair of structural criteria showing what the reasons of detailed balancing are. To formulate this, we need a few concept and also a few formal definitions.

Definition 2. The *circuit conditions* are that the product of reaction rate coefficients along any set of independent cycles is the same in both directions.

Definition 3. Let us take a spanning forest of the Feinberg-Horn-Jackson graph, and let the corresponding reaction step vectors be $\boldsymbol{\gamma}_{.,u}$ $(u = 1, 2, \ldots, N - L)$. Then, $\sum_{u=1}^{N-L} a_u \boldsymbol{\gamma}_{.,u} = \mathbf{0}$ has $N - L - S$ independent solutions. With these $\prod_{u=1}^{N-L} (\frac{k_u^+}{k_u^-})^{a_u} = 1$ should hold: These are the *spanning forest conditions*.

Note that the number of the edges of the spanning *tree* is L less than the number of its vertices (N), if again L is the number of the connected components of the Feinberg-Horn-Jackson graph.

Theorem 2 (Ref. 44). *The mechanism is detailed balanced, if and only if the* circuit conditions *and the* spanning forest conditions *hold.*

An application of Feinberg's theorem (and also the detailed description with examples of the concepts) can be found in Ref. 45.

Example 1. An unconditionally detailed balanced reaction can be seen in Figure 3.

The reason is that both structural conditions are empty:

1. It does not contain cycles.
2. Its deficiency ($:= N - L - S = 3 - 1 - 2$) is zero, thus no spanning forest conditions are to be considered.

With the approach of applying Theorem 1, one sees that $\boldsymbol{\gamma}\mathbf{a} = \mathbf{0}$ has no nonzero solutions, or the kernel of the linear map $\boldsymbol{\gamma}$ only contains the zero vector.

Example 2. A conditionally detailed balanced mechanism is shown in Figure 4. Here the spanning forest conditions (their number is $N - L - S = 5 - 2 - 1 = 2$) are as follows:

$$1 = K_1^2 K_3^{-1} = \frac{k_1^2 k_{-3}}{k_{-1}^2 k_3},\ 1 = K_1 K_2^{-1} = \frac{k_1 k_{-2}}{k_{-1} k_2}. \tag{13}$$

We also have the circuit conditions:

$$1 = K_2 K_3^{-1} K_4 = \frac{k_2 k_{-3} k_4}{k_{-2} k_3 k_{-4}}. \tag{14}$$

Let us try to discuss unconditionally and conditionally detailed balanced reactions in a more systematic way.

If $\boxed{M = 2, R = 2}$, then the following three mass conserving mechanisms remain:

$$\mathrm{X} \underset{k_{-1}}{\overset{k_1}{\rightleftharpoons}} \mathrm{Y}, 2\mathrm{X} \underset{k_{-2}}{\overset{k_2}{\rightleftharpoons}} \mathrm{X} + \mathrm{Y}, \qquad k_{-2} = \frac{k_{-1} k_2}{k_1} \tag{15}$$

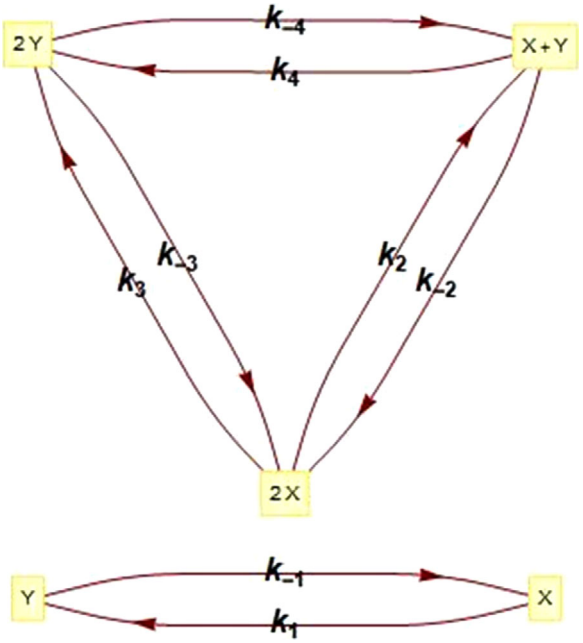

FIGURE 4 Conditionally detailed balanced reaction mechanism: It is detailed balanced if the system of equalities (13) and (14) holds [Color figure can be viewed at wileyonlinelibrary.com]

$$\mathrm{X} \underset{k_{-1}}{\overset{k_1}{\rightleftharpoons}} \mathrm{Y}, 2\mathrm{Y} \underset{k_{-2}}{\overset{k_2}{\rightleftharpoons}} \mathrm{X} + \mathrm{Y}, \qquad k_{-2} = \frac{k_1 k_2}{k_{-1}} \tag{16}$$

$$2\mathrm{X} \underset{k_{-1}}{\overset{k_1}{\rightleftharpoons}} \mathrm{X} + \mathrm{Y}, 2\mathrm{Y} \underset{k_{-2}}{\overset{k_2}{\rightleftharpoons}} \mathrm{X} + \mathrm{Y} \qquad k_{-2} = \frac{k_1 k_2}{k_{-1}}. \tag{17}$$

They are all conditionally detailed balanced. (One can easily show that with two species and two reversible steps there are no unconditionally detailed balanced mechanisms.) They contain no circles, thus only the *spanning forest conditions* should hold, and these are shown in the second column above.

If $\boxed{M = 3, R = 2}$, there are nine conditionally detailed balanced (see Equations 18–26) and 189 unconditionally detailed balanced mechanisms. Let us calculate the number of conditionally and unconditionally detailed balanced mechanisms for the cases $M = 3, 4$. Then, Table 6 results, where MC denotes the total number of mass conserving mechanisms, U is for unconditionally detailed balanced mechanisms, and C stands for conditionally detailed balanced mechanisms: MC = U + C. We use detailed balance as a condition during fitting.

Let us mention that our calculations heavily rely on the program package `ReactionKinetics` written in Wolfram language (*Mathematica*[46]) and downloadable from extras.springer.com using the ISBN number 978-1-4939-8641-5. This package is aimed at helping the chemist to do many kinds of symbolic and numerical investigations of reaction mechanisms including solving the induced kinetic differential equations or simulating the usual stochastic model, but excluding parameter estimation. The codes written to the present paper will be provided to the reader upon request.

TABLE 6 The total number of mechanisms (Tot) and the number of mass conserving (MC) and either unconditionally (U) detailed balanced or conditionally (C) detailed balanced mechanisms with exactly three or four species (M) and one, two, or three reversible reaction steps (R)

M	Number of steps (R)					
	1		2		3	
	Tot	MC = U + C	Tot	MC = U + C	Tot	MC = U + C
3	9	9 = 9 + 0	246	198 = 189 + 9	1994	599 = 0 + 599
4	3	3 = 3 + 0	1302	1254 = 1248 + 6	44 358	28 366 = 24 870 + 3496

3 | APPLICATIONS OF THE METHOD

Instead of demonstrating the method on examples that would require a field-specific knowledge and additional formulations (eg, in combustion: how do temperature and pressure change and how do they affect the rate coefficients), we illustrate it on two simple applications that employ constant rate coefficients and require no other state variables beyond concentrations. Although we use toy models, we think the results are promising. The first generated example shows how noise in "measured" data can impact on model selection, and it picks the best models from nine mass-conserving, conditionally detailed balanced models that can be generated with exactly three species ($M = 3$) and two reversible steps ($R = 2$). The second example is based on real-life data: measurements on the salicylic acid transport, which can be formally treated as a reaction kinetic problem. It shows an example for the inclusion of fictitious species and demonstrates that if the reaction network is not-reversible then at certain initial concentrations an additional filtering criterion may be introduced to drastically reduce the number of candidate mechanisms.

1. *Simulated noisy data.* All nine mass conserving, conditionally detailed balanced reversible mechanisms were constructed with exactly $M = 3$ species and $R = 2$ pairs of reaction steps (see Equations 18–26).

$$\text{M1:}\quad \mathrm{X}\underset{k_{-1}}{\overset{k_1}{\rightleftharpoons}}\mathrm{Y}, \qquad \mathrm{X}+\mathrm{Z}\underset{k_{-2}}{\overset{k_2}{\rightleftharpoons}}\mathrm{Y}+\mathrm{Z} \tag{18}$$

$$\text{M2:}\quad \mathrm{X}\underset{k_{-1}}{\overset{k_1}{\rightleftharpoons}}\mathrm{Z}, \qquad \mathrm{X}+\mathrm{Y}\underset{k_{-2}}{\overset{k_2}{\rightleftharpoons}}\mathrm{Y}+\mathrm{Z} \tag{19}$$

$$\text{M3:}\quad \mathrm{Y}\underset{k_{-1}}{\overset{k_1}{\rightleftharpoons}}\mathrm{Z}, \qquad \mathrm{X}+\mathrm{Y}\underset{k_{-2}}{\overset{k_2}{\rightleftharpoons}}\mathrm{X}+\mathrm{Z} \tag{20}$$

$$\text{M4:}\quad 2\mathrm{X}\underset{k_{-1}}{\overset{k_1}{\rightleftharpoons}}\mathrm{X}+\mathrm{Y}, \qquad \mathrm{X}+\mathrm{Z}\underset{k_{-2}}{\overset{k_2}{\rightleftharpoons}}\mathrm{Y}+\mathrm{Z} \tag{21}$$

$$\text{M5:}\quad 2\mathrm{X}\underset{k_{-1}}{\overset{k_1}{\rightleftharpoons}}\mathrm{X}+\mathrm{Z}, \qquad \mathrm{X}+\mathrm{Y}\underset{k_{-2}}{\overset{k_2}{\rightleftharpoons}}\mathrm{Y}+\mathrm{Z} \tag{22}$$

$$\text{M6:}\quad 2\mathrm{Y}\underset{k_{-1}}{\overset{k_1}{\rightleftharpoons}}\mathrm{X}+\mathrm{Y}, \qquad \mathrm{X}+\mathrm{Z}\underset{k_{-2}}{\overset{k_2}{\rightleftharpoons}}\mathrm{Y}+\mathrm{Z} \tag{23}$$

$$\text{M7:}\quad 2\mathrm{Y}\underset{k_{-1}}{\overset{k_1}{\rightleftharpoons}}\mathrm{Y}+\mathrm{Z}, \qquad \mathrm{X}+\mathrm{Y}\underset{k_{-2}}{\overset{k_2}{\rightleftharpoons}}\mathrm{X}+\mathrm{Z} \tag{24}$$

$$\text{M8:}\quad 2\mathrm{Z}\underset{k_{-1}}{\overset{k_1}{\rightleftharpoons}}\mathrm{X}+\mathrm{Z}, \qquad \mathrm{X}+\mathrm{Y}\underset{k_{-2}}{\overset{k_2}{\rightleftharpoons}}\mathrm{Y}+\mathrm{Z} \tag{25}$$

$$\text{M9:}\quad 2\mathrm{Z}\underset{k_{-1}}{\overset{k_1}{\rightleftharpoons}}\mathrm{Y}+\mathrm{Z}, \qquad \mathrm{X}+\mathrm{Y}\underset{k_{-2}}{\overset{k_2}{\rightleftharpoons}}\mathrm{X}+\mathrm{Z} \tag{26}$$

These are the simplest three-species models with three independent rate coefficients. As both reaction steps within models M1, M2, M3, M4, M5, and M7 describe the same net change of stoichiometric coefficients, they have the same equilibrium constant; thus the detailed balance condition is formally same for them:

$$1 = K_1 K_2^{-1} = \frac{k_1 k_{-2}}{k_{-1} k_2}. \tag{27}$$

Whereas for models M6, M8, and M9, the second step describes a change opposite to that of the first one (ie, $\gamma_{m1} = -\gamma_{m2}, m = 1, 2, 3$), thus the detailed balance condition in their case is different:

$$1 = K_1 K_2 = \frac{k_1 k_2}{k_{-1} k_{-2}}. \tag{28}$$

During fitting, one of these equations was used as a constraint to determine the value of k_{-2} from the other three rate coefficients.

To generate reference data, first, we solved the deterministic model, the induced kinetic differential equations of the first mechanism (Equation 18) with the assumed rate coefficient values:

$$k_1 = 0.1, k_{-1} = 0.1, k_2 = 0.1, \text{thus } k_{-2} = 0.1,$$

which fulfilled the detailed balance condition in Equation 27. We have chosen three sets of initial concentrations, $(x(0), y(0), z(0))$:

$$\{(0.001, 2, 1), (1, 0.001, 2), (2, 1, 0.001)\},$$

which can be obtained from each other by cyclic permutation. Then, we took a sample from the $x(t)$, $y(t)$, and $z(t)$ concentration profiles in the [0,10] time interval at equidistant discrete times with a sampling step size of 0.2,

TABLE 7 Measured concentration values of salicylic acid in the gastric fluid ($x(t)$) and in the intestine fluid ($z(t)$) as a function time

t_i (h)	$x(t_i)$ (mol/dm^3)	t_i (h)	$z(t_i)$ (mol/dm^3)
1	0.01579	0	0
2	0.01429	1	0.0003
3	0.01327	2	0.000614
4	0.01230	3	0.000917
5	0.01148	4	0.00143
6	0.01066	5	0.00201
7	0.00988	6	0.00269
8	0.00912	7	0.00338
9	0.00851	8	0.00402
10	0.00791	9	0.00473

giving data series $(x_{ij}^{\exp}, y_{ij}^{\exp}, z_{ij}^{\exp})$, where $i = 1, 2, 3$ and $j = 0, \ldots, 50$. Finally, to mimic experimental error a normally distributed random noise with 4% and 16% standard deviation was added to each concentration value as a relative error.

The program fitted deterministic kinetic models of all candidate mechanisms separately to both noisy data sets by finding optimum values for the rate coefficients that minimize the following summed square relative deviation objective function:

$$\sum_{i=1}^{3}\sum_{j=0}^{50}\left(\frac{x_{ij}-x_{ij}^{\exp}}{x_{ij}^{\exp}}\right)^2+\left(\frac{y_{ij}-y_{ij}^{\exp}}{y_{ij}^{\exp}}\right)^2+\left(\frac{z_{ij}-z_{ij}^{\exp}}{z_{ij}^{\exp}}\right)^2 \tag{29}$$

and identified the best models based on the Akaike information criterion[47] (AIC).

2. *Real experimental data.* Salicylic acid is an active metabolite of aspirin (acetyl-salicylic acid), which is a common painkiller drug. Its transport in the digestive system: from the gastric fluid (ie, stomach acid, species X) to intestine fluid (species Z) was investigated in a model experiment by measuring its concentration in the two compartments[48] at every whole hour (at the ith hour: $(x_i^{\exp}, z_i^{\exp})$; see Table 7). In this example, no chemical reactions are involved, and the same species in different compartments is denoted with different letters and its transport between compartments is described as formal reactions. The single-step $X \longrightarrow Z$ model did not fit well to the data, and it was found that the concentration changes could be explained by the simple consecutive mechanism: $X \longrightarrow Y \longrightarrow Z$ with the assumption of intermediate fictitious formal species (ie, compartment) Y.

As an application of the presented method, we asked the following question: which of the six kinetic mechanisms $X \xrightarrow{k_1} Y \xrightarrow{k_2} Z$, $X \xrightarrow{k_1} Z \xrightarrow{k_2} Y$, etc, generated by permuting the order of the compartments, can describe the observed concentration changes without having any further information on the nature of the problem. We took fixed intermediate $x_1 = 0.01579$ and initial $y_0 = 0$ and $z_0 = 0$ concentrations (in mol/dm^3 units) and tried to fit all the six models to the data by minimizing the following objective function via tuning rate coefficients (k_1 and k_2) of the two consecutive steps:

$$\sum_{i=2}^{10}(x_i - x_i^{\exp})^2 + \sum_{i=1}^{9}(z_i - z_i^{\exp})^2. \tag{30}$$

The deterministic kinetic models were integrated using an automatic method (function NDSolve in *Mathematica*[46] (version 12.0)), which chooses between Adams, backward differentiation formula, explicit Runge-Kutta, implicit Runge-Kutta, and symplectic partitioned Runge-Kutta methods. The optimal parameters of the candidate models were also determined with an automatic method (function NonlinearModelFit in *Mathematica*), which chooses between the following methods: conjugate gradient, gradient, Levenberg-Marquardt, Newton, Nelder-Mead, differential evolution, simulated annealing, random search, and quasi-Newton. To investigate only physically meaningful models, the square root of rate coefficients was optimized, which, in effect, constrains them to be nonnegative. The optimizations were started with multiple different random initial guess values and ended up in the same minimum, thus we assumed that the global one was found.

3.1 | Simulated noisy data in the $M = 3, R = 2$ case

We fitted the models of the nine reaction mechanisms to the generated data. The estimated values of the parameters with their standard errors (ie, square root of the estimated error variance), root-means-square relative deviation (RMSRD), and the AIC[47] values for fits to data with 4% and 16% standard deviation of the Gaussian noise are shown in Tables 8 and 9, respectively. Rate coefficient values are given with a precision of 10^{-5} as smaller values lead to reaction rates, which cause negligible concentration change on the investigated timescale (ie, 10^1). Figures 5 and 6 show the data and the fitted model solutions for the two noisy data sets. According to the expectations, the original model (model M1) performs well, standard errors of parameters are small, and the original values are within one standard error of the optimal values. Two further models, M4 and M6, are also qualitatively correct and reproduce well the noisy data, whereas the other six models give qualitatively wrong solutions with large deviations for several concentration curves. The relative likelihood of model Mi (eg,

TABLE 8 The rate coefficients and the RMSRD and AIC value for the nine fitted models

M*i*	k_1	k_{-1}	k_2	k_{-2}	RMSRD	AIC
M1	0.09720 (0297)	0.09695 (0349)	0.10317 (0512)	0.10291	0.0484	−1300.3
M2	0.00000 (0000)	0.00000 (0000)	0.04895 (0870)	0.21380	0.4316	529.9
M3	0.07527 (2355)	0.04623 (2062)	0.00006 (2235)	0.00004	0.4532	891.9
M4	0.05493 (0210)	0.05558 (0252)	0.16802 (0455)	0.17001	0.0501	−1245.7
M5	0.00000 (0000)	0.00000 (0000)	0.04894 (0870)	0.21380	0.4316	529.9
M6	0.06699 (0182)	0.07084 (0164)	0.12128 (0401)	0.11469	0.0506	−1265.5
M7	0.02554 (1565)	0.00000 (2755)	0.00000 (0043)	0.00000	0.4301	882.3
M8	0.00000 (0010)	0.00000 (0002)	0.04895 (0870)	0.21380	0.4316	529.9
M9	0.03398 (1098)	0.07685 (1484)	0.00000 (0496)	0.00000	0.4187	856.6

Data were generated by adding a normally distributed relative noise with 4% standard deviation to the simulation results of the reference model. The value of k_{-2} was calculated from the respective detailed balance condition (Equation 27 or Equation 28). Standard error of the parameters is given in brackets as an uncertainty in their last digits.

TABLE 9 The rate coefficients and the RMSRD and AIC value for the nine fitted models

M*i*	k_1	k_{-1}	k_2	k_{-2}	RMSRD	AIC
M1	0.09002 (1090)	0.08916 (01279)	0.11167 (02038)	0.11059	0.1790	−27.7
M2	0.00000 (0006)	0.00000 (00029)	0.04688 (00953)	0.21354	0.4491	636.4
M3	0.06554 (2231)	0.04005 (01988)	0.00000 (02127)	0.00000	0.4751	942.8
M4	0.04951 (0691)	0.04947 (00825)	0.17531 (01775)	0.17520	0.1804	−21.6
M5	0.00000 (0003)	0.00000 (00014)	0.04688 (00953)	0.21354	0.4491	636.4
M6	0.06481 (0677)	0.06855 (00611)	0.12128 (01526)	0.11466	0.1784	−30.8
M7	0.02451 (2318)	0.00000 (03736)	0.00000 (18571)	0.00000	0.4595	937.0
M8	0.12637 (3862)	0.37035 (10882)	0.00000 (00165)	0.00000	0.4491	750.5
M9	0.02709 (0993)	0.06592 (01332)	0.00000 (00550)	0.00000	0.4668	913.0

Data were generated by adding a normally distributed relative noise with 16% standard deviation to the simulation results of the reference model. The value of k_{-2} was calculated from the respective detailed balance condition (Equation 27 or Equation 28). Standard error of the parameters is given in brackets as an uncertainty in their last digits.

M2-M9) with respect to Mj (eg, M1, M6) can be calculated based on their AIC value differences using the following formula:

$$\frac{P(\mathrm{M}i)}{P(\mathrm{M}j)} = \exp\left(-\frac{\mathrm{AIC}_{\mathrm{M}i} - \mathrm{AIC}_{\mathrm{M}j}}{2}\right). \tag{31}$$

According to the AIC differences in Table 8, model M1 is significantly better than M4 and M6 in the case of 4% noise. However, in the case of 16% noise the AIC differences in Table 9 suggest that model M6 is around 5 and 100 times more probable than models M1 and M4, respectively. Consequently, M6 is significantly better (above 95% of confidence) than model M4, whereas this cannot be stated with respect to model M1 (below 95% confidence). Consequently, with the increase of noise (ie, measurement uncertainty) model M6 cannot be distinguished with enough statistical confidence from the correct M1 model and further investigation is needed to decide between them. One possibility to reduce the experimental noise (eg, to 4%) is by applying a better technique or by multiple measurements. Another possibility is to consider other initial conditions that can highlight the differences of similar models.

However, it is not always possible to pick the best model by reducing experimental noise as models can be structurally indistinguishable (see, eg, Ref. 18). Another interesting observation is that models M2 and M5 are effectively identical to the single-step $\mathrm{X} + \mathrm{Y} \underset{k_{-2}}{\overset{k_2}{\rightleftharpoons}} \mathrm{Y} + \mathrm{Z}$ model as their optimal rate coefficient values for the first step (k_1, k_{-1}) are negligibly small, and thereby they provide visually identical simulation results.

3.2 | Salicylic acid transport with $M = 3, R = 2$ case

After brief inspection of the candidate models, it turns out that it is unnecessary to fit all models, because mechanisms $\mathrm{Y} \rightarrow \mathrm{Z} \rightarrow \mathrm{X}$, $\mathrm{Z} \rightarrow \mathrm{X} \rightarrow \mathrm{Y}$, $\mathrm{Z} \rightarrow \mathrm{Y} \rightarrow \mathrm{X}$ provide constant zero concentration for formal species Z, which is known to be definitely different from zero. Furthermore, due to the $y_0 = 0$ assumption, case $\mathrm{Y} \rightarrow \mathrm{X} \rightarrow \mathrm{Z}$ simplifies to the single-step $\mathrm{X} \rightarrow \mathrm{Z}$ mechanism, which is inappropriate.

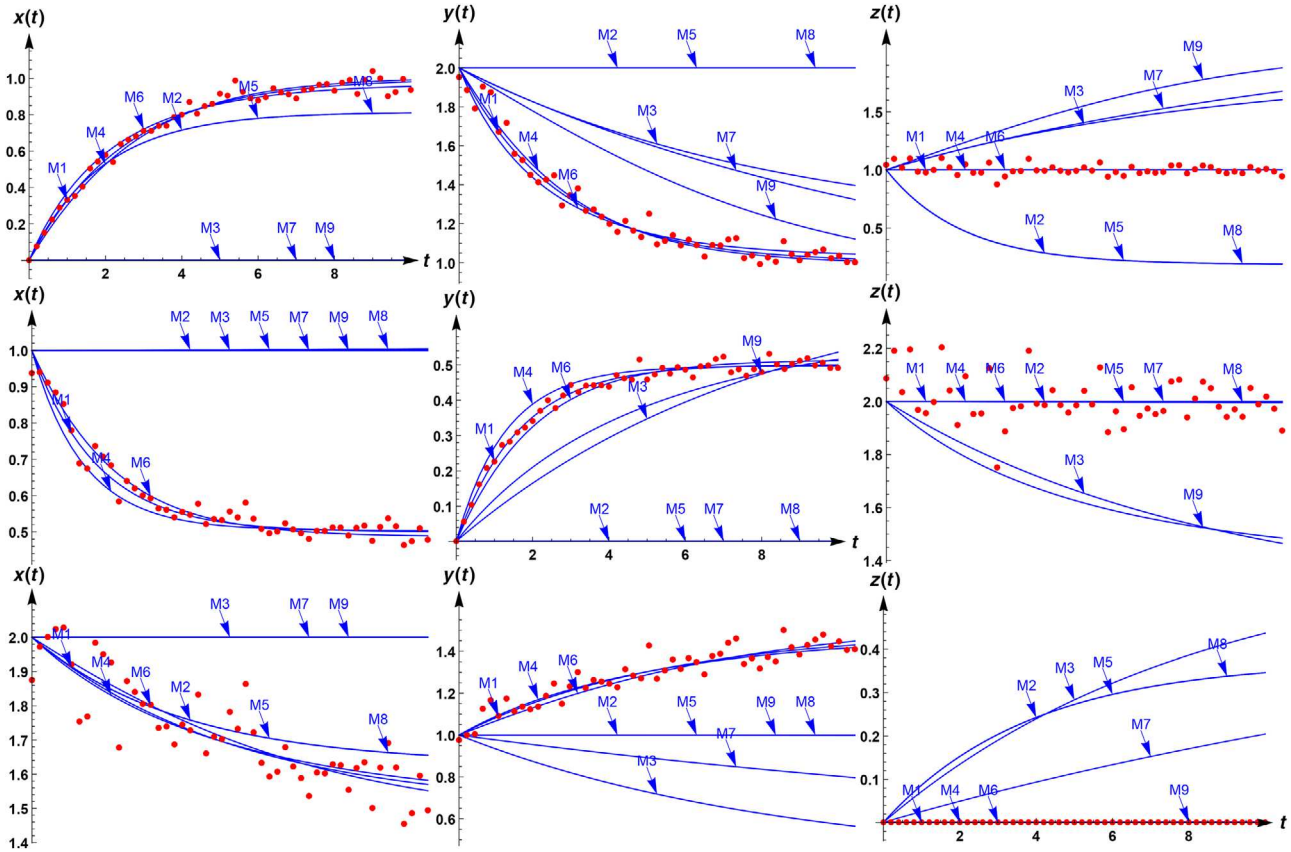

FIGURE 5 Fitting models M1-M9 (blue solid lines; see Equations 18–26) to data (red dots) simulated by model M1 with an added normally distributed relative noise with 4% standard deviation, at three initial conditions shown in rows [Color figure can be viewed at wileyonlinelibrary.com]

How can one systematically filter out these cases in more complex situations? One can use Volpert's theorem[49] as in Ref. 50 to identify those species that will have strictly positive concentrations according to a model for all positive times given a set of species having positive concentrations at zero time. If the species with measured nonzero concentration are not among these species, then the corresponding model cannot be the right one and can be discarded.

Fitting the $X \xrightarrow{k_1} Y \xrightarrow{k_2} Z$ model gives good results: Starting from the initial estimates $(1\,h^{-1}, 1\,h^{-1})$ for the reaction rate coefficients (more precisely, transport coefficients) gives the results with small standard error $\{k_1 = (0.0787 \pm 0.0008)\,h^{-1}, k_2 = (0.181 \pm 0.005)\,h^{-1}\}$, a "good", low correlation value of -0.4997. The fitting can also be considered as good (see the left panel of Figure 7). However, if one picks the wrong $X \rightarrow Z \rightarrow Y$ mechanism, then the concentrations of species Z are fitted badly, as seen in the right panel of Figure 7.

4 | CONCLUSIONS AND OUTLOOK

In this paper, we presented a methodology for kinetic model generation and selection using the framework and theorems of formal reaction kinetics. We demonstrated the method on two formal examples, and thus we only used chemical species without knowing or assuming anything on their elemental composition or structure. One can formulate specific sets of restrictions by taking into consideration the chemical nature of the problem to reduce the number of possible candidate models. From the knowledge of the atomic structure of the species, further restrictions can be derived. For example, explicitly stating the molar mass of the species with measured concentrations reduces the number of mass conserving mechanisms. Furthermore, assuming balance of elements in the reactions, or considering maximum valency of elements when generating fictitious species, constrains the range of possible species and reactions, thereby limits the numbers of them and that of the reaction mechanisms.

We used mass action type kinetics during the integration of the models, and it was also exploited when constraints for conditionally detailed balanced mechanisms was formulated. It is quite natural to ask why we do not select *complex balanced reaction mechanisms*, because this concept turned out to be more fundamental during the development of formal reaction kinetics.[23] Furthermore, in the case of mass action type kinetics we have a nice (structural) necessary and sufficient condition (the reaction mechanism should be weakly reversible

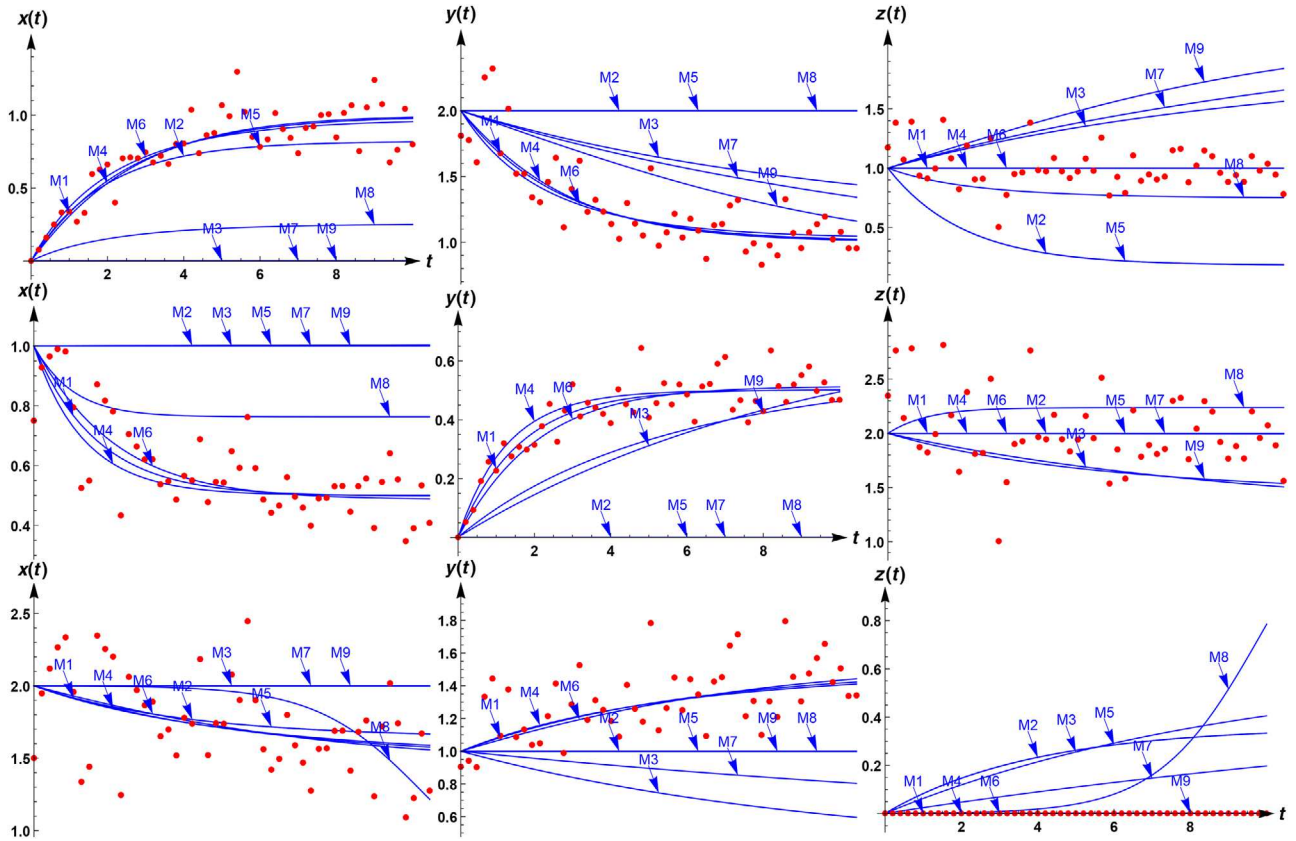

FIGURE 6 Fitting models M1-M9 (blue solid lines; see Equations 18–26) to data (red dots) simulated by model M1 with an added normally distributed relative noise with 16% standard deviation, at three initial conditions shown in rows [Color figure can be viewed at wileyonlinelibrary.com]

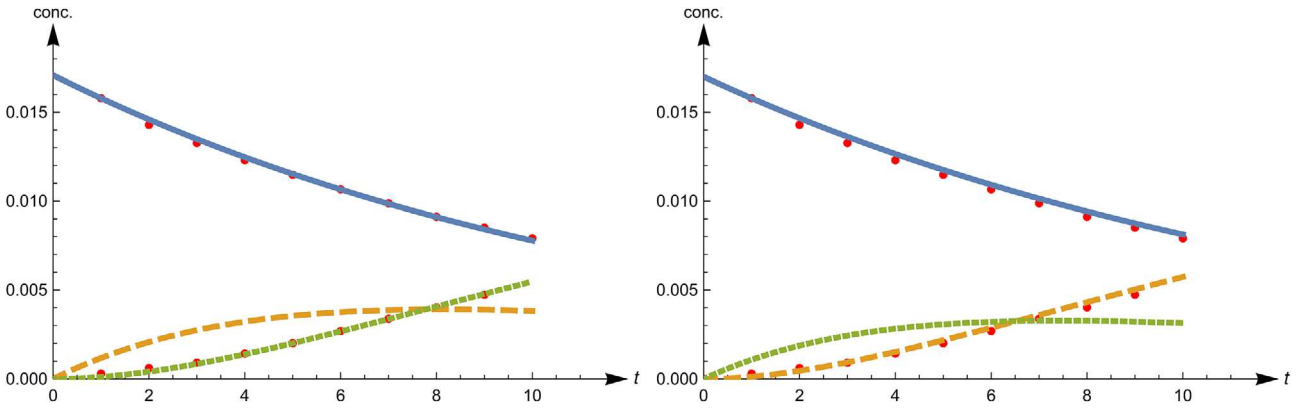

FIGURE 7 Fitting of models $X \xrightarrow{k_1} Y \xrightarrow{k_2} Z$ (left) and $X \xrightarrow{k_1} Z \xrightarrow{k_2} Y$ (right) to experimental data (red dots) on salicylic acid transport between compartments: from stomach (higher values) to intestine (lower values). Salicylic acid in different compartments is represented by different formal species: X (stomach, blue solid line), Z (intestine, green dotted line), and Y (fictitious compartment, orange dashed line) [Color figure can be viewed at wileyonlinelibrary.com]

and should have a zero deficiency $\delta := N - L - S = 0$), this property can also be simply tested. Detailed balancing and complex balancing are not so far from each other as they seem to be.[51–54]

While mechanism generation can be done very quickly, the integration of models multiple times during fitting will take time. Finding *good initial estimates* of the parameters is an important related problem[55] as it can accelerate parameter fitting and help find the global minimum of the objective function, which needs to be done for a large number of candidate models. Once all models have been integrated, we select the best fitting reaction mechanisms using AIC and offer them to

the chemist for further investigations and interpretation. Here we do not want to spend too much time with the detailed statistical analysis of simulated data and the fitted models (see the following books on this topics: Refs. 8,9,16, and 10).

Carefully designed set of restrictions drastically decrease the number of candidate models, and the increasing speed of computers allows the treatment of larger and larger number of candidate models. Depending on the additionally considered restrictions and efficiency improvements in parameter estimation, however, sooner or later the combinatorial blow-up in the number of models wins. Consequently, the proposed procedures are expected to work for systems that can be effectively modeled with not so many real or fictitious species. Our immediate goal is to extend and apply the method to multiple real experimental data sets.

Beyond automatic complete kinetic mechanism generation and selection, the method can be used as a tool for identifying missing reaction steps needed for the interpretation of direct kinetic measurements, which are designed to investigate the rate of a single or very few reactions. Often the assumed steps are not enough to explain the observed concentration profiles. In such case, the proposed method can help to interpret the results by the assumption of additional intermediates or reactions, which may or may not have been postulated. Formally, this situation was demonstrated in the salicylic acid transport example, where the need for an intermediate, Y compartment was identified to explain the measured concentration-time curves. The method presented here can also be applied as a special kind of lumping.[28] Suppose a big kinetic model is given and we construct and parameterize such small models using relevant species (eg, reactants, major products) of the detailed mechanism and some fictitious lumped species in the way described above, that can reproduce the measured or simulated data obtained with the big one. We believe that the proposed method can find application in a wide range of fields where mulitstep kinetic models are applied: in gas kinetics: combustion and atmospheric chemistry, biomass pirolysis,[56,57] liquid phase kinetics,[58] and metabolism (including enzyme kinetics).

The calculations (*Mathematica* notebooks) and data can be requested from the authors.

ACKNOWLEDGMENTS

Irene Otero-Muras raised complex balance as possible property to use for restriction. Judit Zádor has supplied us with (real-life chemical) references continuously. Vilmos Gáspár has carefully read and commented the manuscript. The present work has been supported by the National Research, Development and Innovation Office Hungary (SNN 125739 (JT and TL), PD 120776 (TN)) and the János Bolyai Research Fellowship of the Hungarian Academy of Sciences (BO/00279/16/7 (TN).

ORCID


Tibor Nagy https://orcid.org/0000-0002-1412-3007

REFERENCES

1. Lok L, Brent R. Automatic generation of cellular reaction networks with Moleculizer 1.0. *Nat Biotechnol*. 2005;23(1):131.
2. Sinanoğlu O. Theory of chemical reaction networks. All possible mechanisms or synthetic pathways with given number of reaction steps or species. *J Am Chem Soc*. 1975;97(9):2309-2320.
3. Saunders SM, Jenkin ME, Derwent RG, Pilling MJ. Protocol for the development of the master chemical mechanism, MCM v3 (Part A): tropospheric degradation of non-aromatic volatile organic compounds. *Atmos Chem Phys*. 2003;3(1):161-180.
4. Aumont B, Szopa S, Madronich S. Modelling the evolution of organic carbon during its gas-phase tropospheric oxidation: development of an explicit model based on a self generating approach. *Atmos Chem Phys*. 2005;5(9):2497-2517.
5. Van de Vijver R, Vandewiele NM, Bhoorasingh PL, et al. Automatic mechanism and kinetic model generation for gas- and solution-phase processes: a perspective on best practices, recent advances, and future challenges. *Int J Chem Kinet*. 2015;47(4):199-231.
6. Gao CW, Allen JW, Green WH, West RH. Reaction mechanism generator: automatic construction of chemical kinetic mechanisms. *Comput Phys Commun*. 2016;203:212-225.
7. Proppe J, Reiher M. Mechanism deduction from noisy chemical reaction networks. *J Chem Theory Comput*. 2018;15(1):357-370.
8. Bard Y., *Nonlinear Parameter Estimation*. New York, NY: Academic Press; 1974.
9. Seber GAF, Wild CJ. *Nonlinear Regression*. New York, NY: Wiley; 2003.
10. Weise Th. *Global Optimization Algorithms—Theory and Application*. 3rd ed. E-book. 2011. http://www.it-weise.de/projects/bookNew.pdf, Accessed November 26, 2019.
11. Miller D, Frenklach M. Sensitivity analysis and parameter estimation in dynamic modeling of chemical kinetics. *Int J Chem Kinet*. 1983;15(7):677-696.
12. Frenklach M, Wang H, Rabinowitz MJ. Optimization and analysis of large chemical kinetic mechanisms using the solution mapping method—combustion of methane. *Prog Energy Combust Sci*. 1992;18:47-73.
13. McLean KAP, McAuley KB. Mathematical modelling of chemical processes obtaining the best model predictions and parameter estimates using identifiability and estimability procedures. *Can J Chem Eng*. 2012;90(2):351-366.
14. Sheen DA, Wang H. The method of uncertainty quantification and minimization using polynomial chaos expansions. *Combust Flame*. 2011;158(12):2358-2374.
15. Turányi T, Nagy T, Zsély IGy, et al. Determination of rate parameters based on both direct and indirect measurements. *Int J Chem Kinet*. 2012;44(5):284-302.
16. Turányi T, Tomlin AS. *Analysis of Kinetic Reaction Mechanisms*. Berlin, Germany: Springer; 2014.
17. Varga T, Nagy T, Olm C, et al. Optimization of a hydrogen combustion mechanism using both direct and indirect measurements. *Proc Combust Inst*. 2015;35(1):589-596.
18. Vajda S. Structural equivalence of linear systems and compartmental models. *Math Biosci*. 1981;55(1–2):39-64.

19. Gross E, Harrington HA, Meshkat N, Shiu A. Linear compartmental models: input-output equations and operations that preserve identifiability. *SIAM J Appl Math*. 2019;79(4):1423-1447.
20. Meshkat N, Kuo ChEZ, DiStefano III J On finding and using identifiable parameter combinations in nonlinear dynamic systems biology models and combos: a novel web implementation. *PLoS One*. 2014;9(10): e110261.
21. Sedoglavic A. A probabilistic algorithm to test local algebraic observability in polynomial time. In: *Proceedings of the 2001 International Symposium on Symbolic and Algebraic Computation*. New York, NY: ACM; 2001:309–317.
22. Tóth J, Nagy AL, Papp D. *Reaction Kinetics: Exercises, Programs and Theorems. Mathematica for Deterministic and Stochastic Kinetics*. New York, NY: Springer Nature; 2018.
23. Feinberg M. *Foundations of Chemical Reaction Network Theory*. New York, NY: Springer International Publishing; 2019.
24. Horn F. Stability and complex balancing in mass-action systems with three short complexes. *Proc R Soc Lond A*. 1973;334:331-342.
25. Burke MP, Klippenstein SJ. Ephemeral collision complexes mediate chemically termolecular transformations that affect system chemistry. *Nat Chem*. 2017;9(11):1078-1082.
26. Healy D, Kalitan DM, Aul CJ, Petersen EL, Bourque G, Curran HJ. Oxidation of C1–C5 alkane quinternary natural gas mixtures at high pressures. *Energy Fuels*. 2010;24(3):1521-1528.
27. Nagy T, Turányi T. Reduction of very large reaction mechanisms using methods based on simulation error minimization. *Combust Flame*. 2009;156(2):417-428.
28. Tóth J, Li G, Rabitz H, Tomlin AS. The effect of lumping and expanding on kinetic differential equations. *SIAM J Appl Math*. 1997;57:1531-1556.
29. Frank-Kamenetskii DA. *Diffusion and Heat Transfer in Chemical Kinetics* [in Russian]. Moscow, USSR: USSR Academy of Science Press; 1947.
30. Einstein A. *Relativity*. London: Routledge; 2013.
31. Horie O, Crowley JN, Moortgat GK. Methylperoxy self-reaction: products and branching ratio between 223 and 333 K. *J Phys Chem*. 1990;94(21):8198-8203.
32. Noell AC, Alconcel LS, Robichaud DJ, Okumura M, Sander SP. Near-infrared kinetic spectroscopy of the HO_2 and $C_2H_5O_2$ self-reactions and cross reactions. *J Phys Chem A*. 2010;114(26):6983-6995.
33. Sloane NJA, ed. *The On-Line Encyclopedia of Integer Sequences*. https://oeis.org. Accessed November 26, 2019.
34. Lovász L. *Combinatorial Problems and Exercises*. Providence, RI: AMS Chelsea Publishing; 2007.
35. Maxwell JC. IV. On the dynamical theory of gases. *Philos Trans R Soc Lond*. 1867;157:49-88.
36. Boltzmann L. *Lectures on Gas Theory*. Berkeley, CA: University of California Press; 1964.
37. Einstein A. Strahlungs-Emission und -Absorption nach der Quantentheorie. *Verh Dtsch Phys Ges*. 1916;18:318-323.
38. Wegscheider R. Über simultane Gleichgewichte und die Beziehungen zwischen Thermodynamik und Reaktionskinetik homogener Systeme. *Z Phys Chem*. 1901/2;39:257-303.
39. Horn F, Jackson R. General mass action kinetics. *Arch Ration Mech Anal*. 1972;47:81-116.
40. Fowler RH, Milne EA. A note on the principle of detailed balancing. *Proc Natl Acad Sci USA*. 1925;11:400-402.
41. Tolman RC. The principle of microscopic reversibility. *Proc Natl Acad Sci USA*. 1925;11:436-439.
42. Gold V, Loening KL, McNaught AD, Shemi P. *IUPAC Compendium of Chemical Terminology*. 2nd ed. Oxford, England: Blackwell Science; 1997.
43. Vlad MO, Ross J. Thermodynamically based constraints for rate coefficients of large biochemical networks. *Syst Biol Med*. 2009;1(3):348-358.
44. Feinberg M. Necessary and sufficient conditions for detailed balancing in mass action systems of arbitrary complexity. *Chem Eng Sci*. 1989;44(9):1819-1827.
45. Nagy I, Tóth J. Microscopic reversibility or detailed balance in ion channel models. *J Math Chem*. 2012;50(5):1179-1199.
46. Wolfram Research, Inc. *Mathematica, Version 12.0*. Champaign, IL: Wolfram Research; 2019.
47. Akaike H. A new look at the statistical model identification. *IEEE Trans Automat Contr*. 1974;19(6):716-723.
48. Rácz I, Gyarmati I, Tóth J. Effect of hydrophilic and lipophilic surfactant materials on the salicylic acid transport in a three compartment model [in Hungarian]. *Acta Pharm Hung*. 1977;47:201-208.
49. Volpert AI, Hudyaev SI. *Analyses in Classes of Discontinuous Functions and Equations of Mathematical Physics*. Dordrecht, the Netherlands: Martinus Nijhof; 1985. original, 1975 [in Russian].
50. Kovács K, Vizvári B, Riedel M, Tóth J. Decomposition of the permanganate/oxalic acid overall reaction to elementary steps based on integer programming theory. *Phys Chem Chem Phys*. 2004;6(6):1236-1242.
51. Dickenstein A, Millán MP. How far is complex balancing from detailed balancing? *Bull Math Biol*. 2011;73:811-828.
52. Gorban AN, Yablonsky GS. Extended detailed balance for systems with irreversible reactions. *Chem Eng Sci*. 2011;66(21):5388-5399.
53. van der Schaft A, Rao Sh, Jayawardhana B. Complex and detailed balancing of chemical reaction networks revisited. *J Math Chem*. 2015;53(6):1445-1458.
54. Szederkényi G, Hangos KM. Finding complex balanced and detailed balanced realizations of chemical reaction networks. *J Math Chem*. 2011;49:1163-1179.
55. Ladics T, Tóth J, Nagy T. Automatic initial estimates. *Int J Chem Kinet*. 2020. (in preparation).
56. Antal MJ, Várhegyi G. Cellulose pyrolysis kinetics—the current state knowledge. *Ind Eng Chem Res*. 1995;34(3):703-717.
57. Dussan K, Dooley S, Monaghan RFD. A model of the chemical composition and pyrolysis kinetics of lignin. *Proc Combust Inst*. 2019;37(3):2697-2704.
58. Csekő G, Valkai L, Horváth AK. A simple kinetic model for description of the iodate–arsenous acid reaction: experimental evidence of the direct reaction. *J Phys Chem A*. 2015;119(45):11053-11058.

How to cite this article: Nagy T, Tóth J, Ladics T. Automatic kinetic model generation and selection based on concentration versus time curves. *Int J Chem Kinet*. 2020;52:109–123. https://doi.org/10.1002/kin.21335